Архитектура нейронный сетей: Трансформеры

Введение в курс

Содержание курса

1. Motivation for Transformer. Attention. The original Transformer architecture.
2. Transformer-based Encoders. Masked language models based on the Transformer architecture. BERT and related models.
3. Classification and sequence tagging with Transformers. Using encoders to generate feature representation for various NLU tasks.
4. Transformer-based Decoders. Generation of text based on the Transformer architecture. GPT and related decoders. Text generation methods. Prompt tuning.
5. Prompt and Instruction tuning. Reinforcement Learning from Human Feedback (RLHF), ChatGPT, and related models.
6. Sequence to sequence tasks: machine translation, text detoxification, question answering, dialogue. Technical tricks for training and inference: infrastructure and performance.
7. Multilingual language models based on the Transformer architecture.
8. Efficient Transformers.
9. Compression of transformer models.
10. Network encoders with Transformers.
11. Multimodal and vision Transformers.
12. Transformers for tabular data.
13. Transformers for event sequences.

Преподаватели

Александр Панченко	Разработка курса в целом
Ирина Никишина	Домашние задания
Алексей Зайцев	
Мария Тихонова	
Виктория Чекалина	
Антон Разжигаев	

Формат курса

- 13 занятий по вторникам и четвергам, 18:00 - 20:20
  - Перерыв 10 мин между лекцией (1ч) и семинаром (1ч)
  - 10-15 мин на вопросы после каждой сессии
  - Читаем на русском, но слайды преимущественно **на английском**
- Квизы по лекциям (на англ, оцениваются автоматически)
- Задания по семинарам (1 задание на доп. эксперименты)
- Две домашние работы
- Финальный квиз

Формула оценки: 40%ДЗ1 + 40%ДЗ2 + 20%Тест (+ бонус)

Бонус = 14 баллов за квизы + 14 за задания по семинару

Почему рабочий язык – английский

- Подавляющая часть существующих материалов – на английском
- Терминология – на английском
- Статьи – на английском
- Основная терминология – транслитерация с английского (промптинг, трансформер, файнтьюнинг)

Мы обязательно будем помогать с терминологией и переводом, обязательно спрашивайте, если что-то непонятно!

Рекомендуемая литература

- Dan Jurafsky and James H. Martin (2021). Speech and Language Processing (3rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/
- Denis Rothman (2021). Transformers for Natural Language Processing. https://www.packtpub.com/product/transformers-for-natural-language-processing/9781800565791
- The Hugging Face NLP course. https://huggingface.co/course

00. Basic neural architectures

Уже в системе:

- Вводная лекция по основам нейросетевых структур
- Семинар по pytorch и основам нейросетевых структур
- Квиз по лекционной части
- Задание к семинару

Домашние задания

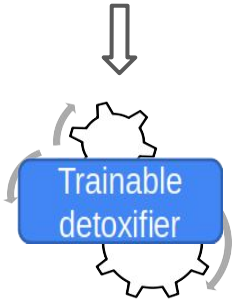
Assignments

Semantic Role labeling

Postgres	B-OBJECT
is	O
easier	B-PREDICATE
to	O
install	B-ASPECT
than	O
Microsoft	B-OBJECT
SQL	I-OBJECT

RUSSE'2022 Detoxification

> You idiot, stop talking of what you have no idea about!



> I would suggest learning more about the subject to get a more productive conversation.

Semantic role tagging

RUSSE-2022 Detoxification

DIALOGUE 2022

Dialogue Evaluation

Russian Text Detoxification Based on Parallel Corpora

You are very welcome to the first shared task on text detoxification based on a parallel dataset!

https://www.dialog-21.ru/en/evaluation/

https://russe.nlpub.org/2022/tox/

RUSSE-2022 Detoxification

	toxic sentence	**detoxified sentence**
	из за таких [redacted] мы и страдаем	Из-за таких людей мы и страдаем
translation:	*We suffer from such [redacted]*	*We suffer from such people*
	[redacted] знает кто кум, но девушка красивая 👍	неизвестно кто кум, но девушка красивая
translation:	*[redacted] knows who the godfather is, but the girl is beautiful* 👍	*it is unknown who the godfather is, but the girl is beautiful*
	порядок бы [redacted] навёл !	Порядок бы навел
translation:	*Put these [redacted] things in order*	*Put the things in order*

Система оценивания

Технический отчет		**Код**		**Результаты**		**Сумма баллов**
Методология	Анализ результатов	Читаемость	Воспроизводимость	Преодоление базового решения	Top-1 +10 баллов Top-20% +5 баллов	100% + бонус
5	5	5	5	0 / 5	0 / 5 / 10	25 (max 35)

Технический отчет

[Colab link]

2. Technical Report

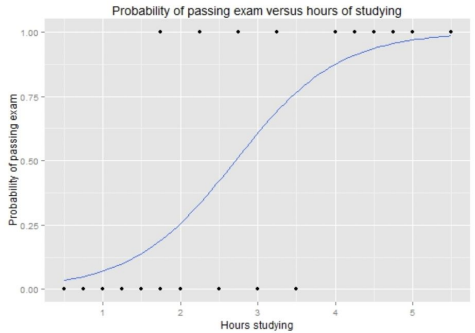
Use Section 2 to describe results of your experiments as you would do writing a paper about your results. DO NOT insert code in this part. Only insert plots and tables summarizing results as needed. Use formulas if needed do described your methodology. The code is provided in Section 3.

2.1 Methodology

***Enter here** a detailed description of the method used in your submission(s) to Codalab. The description should be at least 2-4 paragraphs featuring the following: type of the model, meta-parameters, how did you select meta-parameters, any further modifications of the out-of-the-box solutions, etc. The text is markdown and you can use math environment to write formulas:*

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} x_j \beta_j$$

Also you can insert images as needed:



This part of the should contain description of all methods that you tried and, most importantly, that worked the best for you. Here you can include some tricks of your preprocessing, description of the models and motivation of their usage, the description of the training process details (train-test split, cross-validation, etc.). So, everything valuable that will help us to understand the scope of your work and reproduce

Технический отчет

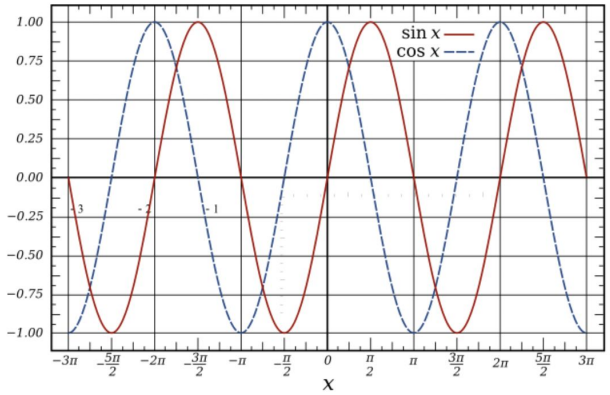
[Colab link]

2.2 Discussion of results

***Enter here** a discussion of results and a summary of the experiment. Here we want to see the final table with comparison of the baseline and all tried approaches you decided to report. Even if some method did not bring you to the top of the leaderboard, you should nevertheless indicate this result and a discussion, why, in your opinion, some approach worked and another failed. Interesting findings in the discussion will be a plus.*

Method	Precision	Recall
Baseline	0.88	0.77
My great method 1	0.99	0.11
My great method 2	0.90	0.90

*If relevant insert plots and historgams in this section e.g. testing variation of the score with respect to some parameters e.g. learning rate or size of the input dataset, etc. Please do not use code to generate plots, instead just insert images as shown below. Plots could be generated from code in Section 3. *



Код

3. Code

Enter here all code used to produce your results submitted to Codalab. Add some comments and subsections to navigate though your solution.

In this part you are expected to develop yourself a solution of the task and provide a reproducible code:

- *Using Python 3;*
- *Contains code for installation of all dependencies;*
- *Contains code for downloading of all the datasets used;*
- *Contains the code for reproducing your results (in other words, if a tester downloads your notebook she should be able to run cell-by-cell the code and obtain your experimental results as described in the methodology section).*

As a result, you code will be graded according to these criteria:

- ***Readability**: your code should be well-structured preferably with indicated parts of your approach (Preprocessing, Model training, Evaluation, etc.).**
- ***Reproducibility**: your code should be reproduced without any mistakes with "Run all" mode (obtaining experimental part).**

3.1 Requirements

[ ] ↳ *1 cell hidden*

3.2 Download the data

[ ] ↳ *2 cells hidden*

3.3 Preprocessing

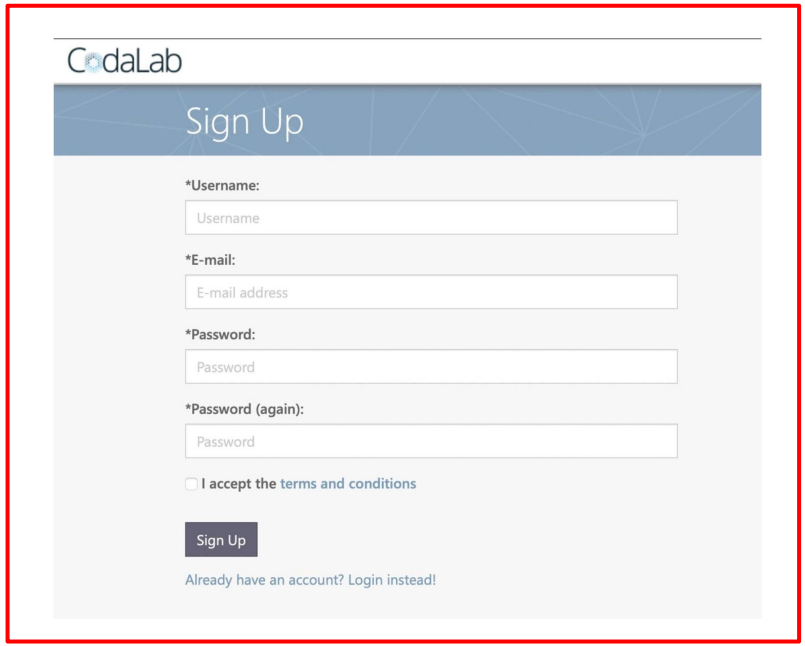
[ ] ↳ *1 cell hidden*

3.4 My method of text processing

↳ *1 cell hidden*

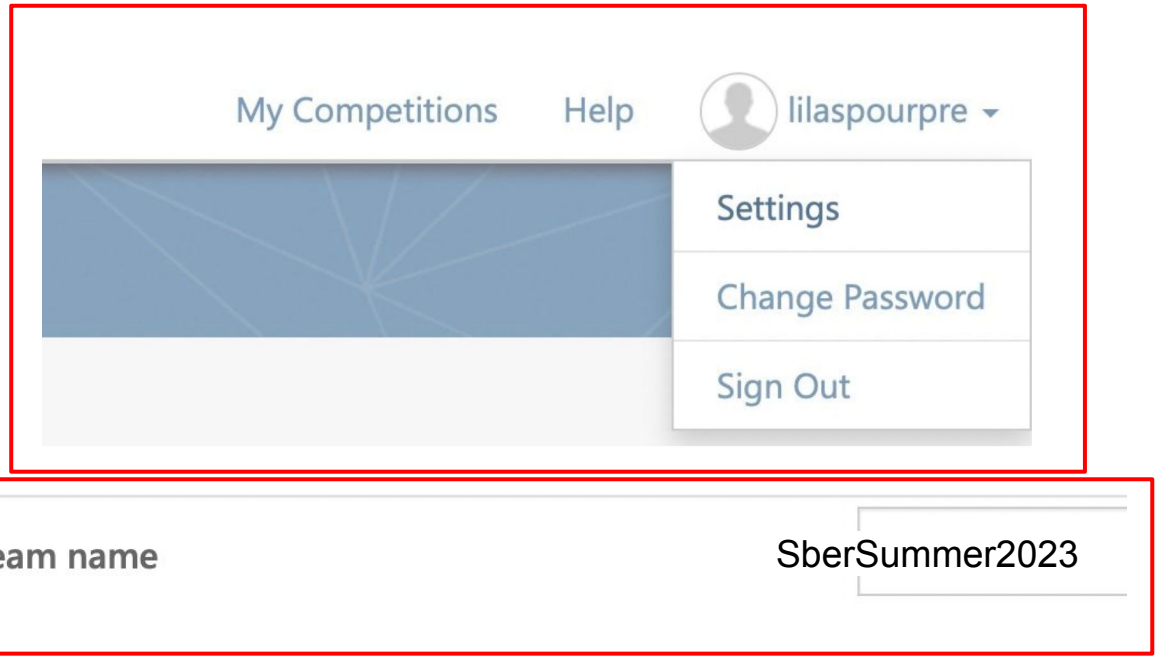
CodaLab (регистрация)

Semantic role labelling: https://codalab.lisn.upsaclay.fr/competitions/531
Detoxification: https://codalab.lisn.upsaclay.fr/competitions/642

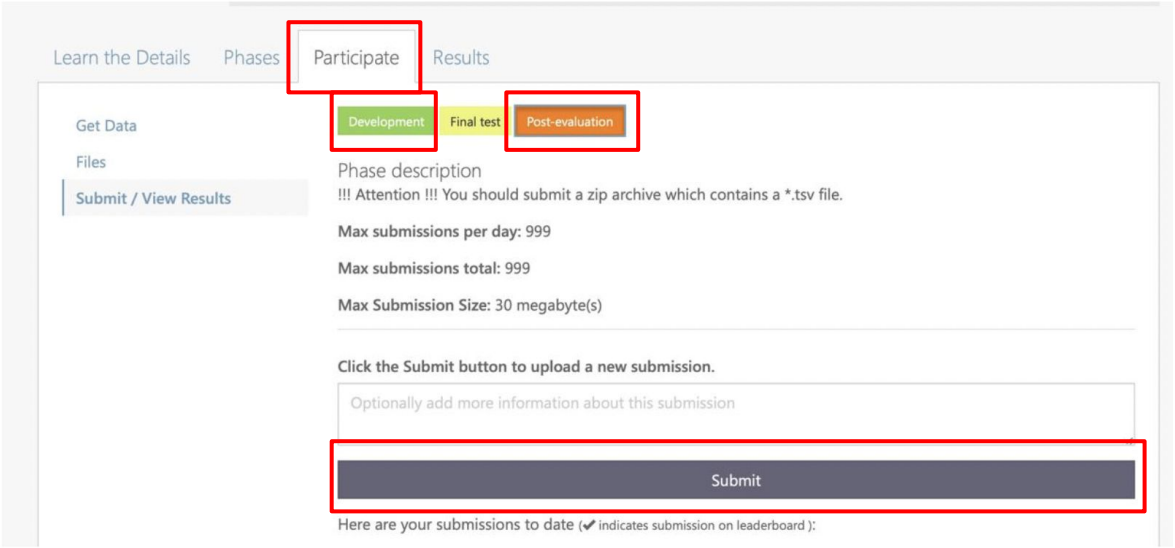


CodaLab (необходимо указать имя команды)

Semantic role labelling: https://codalab.lisn.upsaclay.fr/competitions/531
Detoxification: https://codalab.lisn.upsaclay.fr/competitions/642



Отправка решения



Лидерборд

Learn the Details | Phases | Participate | Results

Development | Final test | Post-evaluation

Phase description

!!! Attention !!! You should submit a zip archive which contains a *.txt file with your detoxified sentences.

Max submissions per day: 999

Max submissions total: 100

Download CSV | Download all submissions on leaderboard

Results

#	User	Entries	Date of Last Entry	Team Name	Style transfer accuracy ▲	Meaning preservation ▲	Fluency ▲	Joint score ▲	ChrF1 ▲
1	gleb.shnshn	3	02/04/22		0.97 (3)	0.93 (1)	0.95 (4)	0.86 (1)	0.53 (13)
2	orzhan	24	01/25/22		0.98 (2)	0.83 (6)	0.95 (2)	0.78 (2)	0.55 (10)
3	Phoenix120	24	01/23/22	Mindful Squirrel	0.91 (7)	0.90 (2)	0.95 (1)	0.78 (3)	0.55 (8)
4	Rexhaif	1	01/03/22	FRC CSC RAS	0.93 (5)	0.83 (7)	0.95 (3)	0.73 (4)	0.57 (4)
5	aorews	1	02/08/22		0.94 (4)	0.79 (13)	0.85 (9)	0.66 (5)	0.52 (15)

Deadline: 11.08